

1 Claims:

2
3 1. A collection of software tools for acquiring data from
4 diverse sources and/or structuring the data and/or determining
5 similarity of content, said collection comprising:

6 one or more tools selected from the group consisting of a
7 web agent creator, a web agent created by the web agent creator,
8 a web agent manager, an ontology-directed classifier, an
9 ontology-directed extractor, and an ontology-directed matcher.

10
11 2. The collection according to claim 1, wherein:
12 one or more of the tools are example driven through a
13 graphical user interface.

14
15 3. The collection according to claim 1, wherein:
16 said web agent creator has a web browser interface and a
17 web agent is created by navigating to a web page of interest and
18 selecting the kind of information to be extracted from the web
19 page.

1 4. The collection according to claim 1, wherein:

2 said web agent creator includes

3 a web browser user interface,

4 a pattern expression discovery algorithm coupled to

5 said user interface,

6 a results editor coupled to said user interface and

7 said pattern expression discovery algorithm,

8 an agent generator coupled to said user interface and

9 said results editor, and

10 a form value editor coupled to said user interface and

11 said agent generator.

12
13 5. The collection of claim 4, wherein:

14 said user interface indicates text selected by the user

15 interface to said pattern expression discovery algorithm, said

16 results editor, said agent generator, and said form value

17 editor.

6. The collection of claim 4, wherein:

said pattern expression discovery algorithm is an XPath
discovery algorithm,
said user interface indicates a DOM tree of text selected
by the user interface to said XPath discovery algorithm, said
results editor, said agent generator , and said form value
editor.

7. The collection of claim 5, wherein:

said pattern expression discovery algorithm generates a
pattern expression based on the results received from the user
interface and communicates that pattern expression to the
results editor.

8. The collection of claim 6, wherein:

said XPath discovery algorithm generates an XPath based on
the DOM tree received from the user interface and communicates
that XPath to the results editor.

9. The collection of claim 7, wherein:

the results editor receives pattern expressions from the
pattern expression discovery algorithm and accepts input from
the user interface to identify the nature of the selected text.

1 10. The collection of claim 8, wherein:

2 the results editor receives XPath expressions from the
3 XPath discovery algorithm and accepts input from the user
4 interface to identify the nature of the selected text.

5
6 11. The collection of claim 8, wherein:

7 the form value editor receives input from the user
8 interface and provides output to the agent generator including
9 instructions and data to be used by the agent generated by the

10 agent generator to fill out web based forms in order to reach
11 the source of data to be extracted by the agent.

12
13 12. The collection of claim 11, wherein:

14 the pattern expression discovery algorithm takes as its
15 input a set of items corresponding to the text highlighted by
16 the user interface,
17 identifies the items, and
18 determines corresponding data extractor and isolator
19 expressions.

1 13. The collection of claim 11, wherein:

2 the pattern expression discovery algorithm is an XPath

3 discovery algorithm,

4 the XPath discovery algorithm takes as its input a set of

5 nodes corresponding to the text highlighted by the user

6 interface,

7 identifies locator nodes and grouping nodes based on the

8 input set of nodes, and

9 determines corresponding data extractor and isolator

10 expressions.

11
12 14. The collection according to claim 12, wherein:

13 the corresponding data extractor and isolator expressions

14 are used to form a navigation map to be used by the agent to

15 find all nodes that match the isolator expression, and

16 for each node matching the isolator expression, find a

17 match for each of the data extractor expressions.

18
19 15. The collection according to Claim 1, wherein:

20 the ontology directed classifier uses a taxonomy provided

21 by a tree of classes and subclasses generated using an ontology

22 management system.

1 16. The collection according to Claim 15, wherein:

2 the ontology directed classifier performs taxonomy token
3 weighting, node weighting for descriptions, weight propagation
4 and normalizations, and determining the best class and subtree
5 of said taxonomy to which an item can be classified.

6
7 17. The collection according to claim 1, wherein:

8 said ontology directed extractor takes unstructured text
9 descriptions about an item as input and produces a set of
10 structured property values about the item as output.

11
12 18. A web agent creator for creating a web agent to
13 acquire data from the world wide web, said web agent creator
14 comprising:

15 a web browser user interface,
16 a pattern expression discovery algorithm coupled to
17 said user interface,
18 a results editor coupled to said user interface and
19 said pattern expression discovery algorithm,
20 an agent generator coupled to said user interface and
21 said results editor, and
22 a form value editor coupled to said user interface and
23 said agent generator.

1 19. The web agent creator according to claim 18, wherein:
2 said user interface indicates text selected by the user
3 interface to said pattern expression discovery algorithm, said
4 results editor, said agent generator, and said form value
5 editor.

6
7 20. The web agent creator according to claim 18, wherein:
8 said pattern expression discovery algorithm is an XPath
9 discovery algorithm,
10 said user interface indicates a DOM tree of text selected
11 by the user interface to said XPath discovery algorithm, said
12 results editor, said agent generator , and said form value
13 editor.

14
15 21. The web agent creator according to claim 19, wherein:
16 said pattern expression discovery algorithm generates a
17 pattern expression based on the results received from the user
18 interface and communicates that pattern expression to the
19 results editor.

20
21 22. The web agent creator according to claim 20, wherein:
22 said XPath discovery algorithm generates an XPath based on
23 the DOM tree received from the user interface and communicates
24 that XPath to the results editor.

1 23. The web agent creator according to claim 18, wherein:
2 the results editor receives pattern expressions from the
3 pattern expression discovery algorithm and accepts input from
4 the user interface to identify the nature of the selected text.

5
6 24. The web agent creator according to claim 20,
7 wherein:the results editor receives XPath expressions from the
8 XPath discovery algorithm and accepts input from the user
9 interface to identify the nature of the selected text.

10
11 25. The web agent creator according to claim 18, wherein:
12 the form value editor receives input from the user
13 interface and provides output to the agent generator including
14 instructions and data to be used by the agent generated by the
15 agent generator to fill out web based forms in order to reach
16 the source of data to be extracted by the agent.

17
18 26. The web agent creator according to claim 18, wherein:
19 the pattern expression discovery algorithm takes as its
20 input a set of items corresponding to the text highlighted by
21 the user interface,
22 identifies the items, and
23 determines corresponding data extractor and isolator
24 expressions.

1 27. The web agent creator according to claim 18, wherein:
2 the pattern expression discovery algorithm is an XPath
3 discovery algorithm,
4 the XPath discovery algorithm takes as its input a set of
5 nodes corresponding to the text highlighted by the user
6 interface,
7 identifies locator nodes and grouping nodes based on the
8 input set of nodes, and
9 determines corresponding data extractor and isolator
10 expressions.

11

12 28. The web agent creator according to claim 26, wherein
13 the corresponding data extractor and isolator expressions
14 are used to form a navigation map to be used by the agent to
15 find all nodes that match the isolator expression, and
16 for each node matching the isolator expression, find a
17 match for each of the data extractor expressions.

18

19 29. An ontology directed classifier for use with an
20 ontology management system, said ontology directed classifier
21 comprising:

22 means for receiving a taxonomy as input; and

23 means for generating a tree of classes and subclasses as
24 output for use by the ontology management system.

1

2 30. The ontology directed classifier according to claim 29,

3 further comprising:

4 means for taxonomy token weighting,

5 means for node weighting for descriptors

6 means for weight propagation and normalization, and

7 means for determining the best class and sub-tree of said

8 taxonomy to which an item can be classified.

9

10 31. An ontology directed extractor for use with an

11 ontology management system, said ontology directed extractor,

12 comprising:

13 means for receiving an unstructured text description about

14 an item as input, and

15 means for producing a set of structured property values

16 about the item as output.

17

18 32. An ontology directed extractor according to claim 31,

19 wherein:

20 said structured property values are structured by ontology

21 relationships.

1 33. An ontology directed matcher for use with an ontology
2 management system, said ontology directed matcher comprising:

3 means for describing items based on a structured set of
4 properties;

5 means for defining the relative importance of said
6 properties in describing said items; and

7 means for scoring the degree of equivalence of items based
8 on said definitions
9

10 34. An ontology directed matcher according to claim 33,
11 wherein:

12 said structured set of properties in defined by ontology
13 attributes provided by the ontology management system.
14

15 35. An ontology directed matcher according to claim 34,
16 wherein:

17 said means for defining the relative importance of said
18 properties is based on weight attached to a matching function
19 for each said property that takes as input the values of said
20 attributes defining that property for two different items and
21 outputs a number indicating the similarity of these input
22 values.
23
24

1

2 36. An ontology directed matcher according to claim 35,

3 wherein:

4 said means for scoring the degree of equivalence of items

5 includes means for multiplying the said output values of all

6 said matching functions by said respective weights and summing

7 these products.

8

9 37. The collection according to claim 1, further

10 comprising:

11 a validation method applied to one or more tools in the

12 collection to determine the accuracy of the tool's output by

13 manually checking the accuracy of a statistical sampling of tool

14 output from specific tool input.

15

16 38. The collection according to claim 37, wherein:

17 said validation method determines an Acceptable Quality

18 Level (AQL) as defined in standard ANSI/ASQC Z1.4-1993 by

19 performing multiple sampling procedures at different AQLs as

20 defined in said standard until the boundary AQL level is found

21 below which the sampling procedure fails and above which the

22 sampling procedure succeeds.